Chapter 2

Complex Linkages Made Easy

John R. H. Charlton, Office for National Statistics, UK
Judith D. Charlton, JDC Applications

Abstract

Once valid key fields have been set up, relational database techniques enable complex linkages that facilitate a number of statistical analyses. Using one particular example, a classification of types of linkages is developed and illustrated. The naive user of such data would not necessarily know how to use a relational database to perform the linkages, but may only know the sort of questions they want to ask. To make data (anonymous to protect the confidentiality of patients and doctors) generally accessible, a user-friendly front-end has been written using the above concepts, which provides flat-file datasets (tabular or list) in response to answers from a series of questions. These datasets can be exported in a variety of standard formats. The software will be demonstrated, using a sample of the data.

Introduction

Most of the papers at this workshop are concerned with establishing whether or not different records in the database match. This paper starts from the point where this matching has already been established. It will thus be assumed that the data have already been cleaned, duplicates eliminated, and keys constructed through which linkages can be made. Procedures for matching records when linkage is not certain have been discussed for example by Newcombe *et al.* (1959, 1988), Fellegi and Sunter (1969), and Winkler (1994). We also assume database software that can:

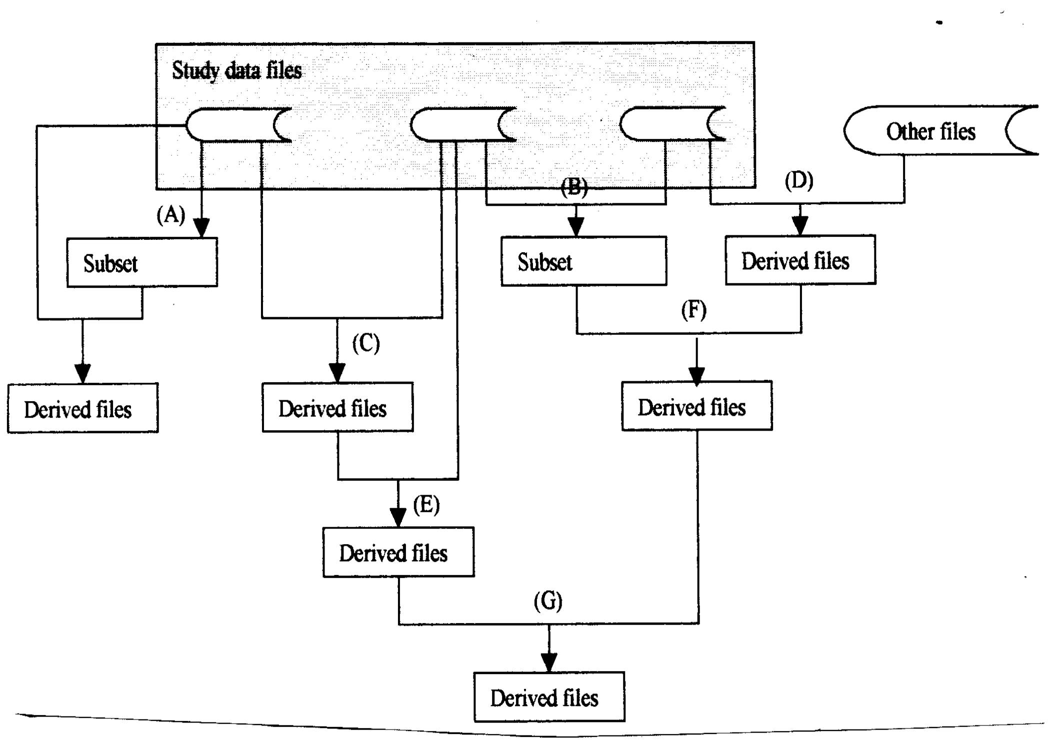
- select fields from a file of records;
- extract from a file either;
  - all records
  - distinct records which satisfy specified criteria; and
- link files using appropriate key and foreign fields.

The purpose of this paper is firstly to illustrate the huge potential of using relational databases for data linkage for statistical analyses. In the process a classification of linkages will be developed, using a particular database to illustrate the points. Some results will be presented by way of example. We will show how the complex linkages required for statistical analyses can be decomposed into a sequence of simple database queries and linkages. Finally a user-friendly program that has been written for extracting a number of different types of dataset for analyses will be described. The advantages and disadvantages of such approaches will be discussed.

Relational databases are ideal for storing statistical data, since they retain the original linkages in the data, and hence the full data structure. They also facilitate linking in new data from other sources, and are economical in data storage requirements. However, most statistical analyses require simple rectangular files, and complex database queries may be required to obtain these.

The linkages required to obtain the flat files for statistical analysis vary from the relatively simple to the extremely complex (Figure 1). Subsets of data files may be found (A), possibly by linkage with another file (B). Derived files may be created by linking files (or their subsets) within the study data files (C), or to files outside the study data (D). The derived files may be further linked to files in or outside the dataset (E), subsets (F), or other derived files (G), to obtain further derived files, and this process may continue at length.

Figure 1. – Illustration of Complex Linkage

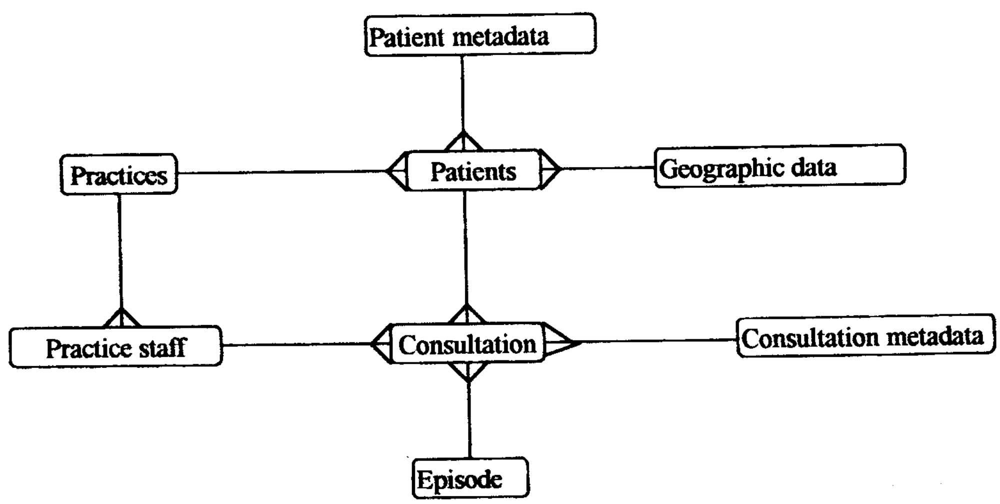


The Example Database

In a major survey in England and Wales (MSGP4) some 300 general medical practitioners (GPs) in 60 practices collected data from half a million patients, relating to every face to face contact with them over the course of a year (McCormick *et al.*, 1995). In the UK nearly the entire population is registered with a GP, and only visit a doctor in the practice in which they are registered, except in an emergency, when they may attend an accident and emergency department of a hospital or another GP as a temporary patient. For all patients in MSGP4 there was information on age, sex, and postcode. In addition socio-economic data were successfully collected by interview for 83 per cent of the patients on these doctors registers. There was a core of common questions, but there were also questions specific to children, adults, and married/cohabiting women. Information was also collected about the practices (but not individual GPs). Geographic information related to postcodes was also available. The structure of the data is illustrated in

Figure 2 (simplified). MSGP4 was the fourth survey of morbidity in general practice. In previous MSGP surveys output consisted only as a series of tables produced by COBOL programs, and MSGP4 was the fist survey for which relational databases were used to provide flexible outputs.

Figure 2.—Example Database (Simplified) – Patient Consultations of General Medical Practitioners



Some Definitions

- **Read Code**.--A code used in England and Wales by general practice staff to identify uniquely a medical term. This coding was used in the MSGP project because it is familiar to general practice staff, but it is not internationally recognised and the codes have a structure that does not facilitate verification.

- **ICD Code**.--International classification of disease. Groups of Read codes can be mapped onto ICD9 codes. For example Read code F3810= "acute serious otitis media," maps to ICD A381.0 = "acute nonsuppurative otitis media"). Such mappings form part of the consultation metadata (see below).

- **Consultation.--**A "consultation" refers to a particular diagnosis by a particular member of staff on a particular date at a particular location, resulting from a face to face meeting between a patient and doctor/ nurse. A "diagnosis" is identified by a single Read code.

- **"Patients Consulting"**.--Some registered patients did not consult a doctor or other staff member during the study. "Patients consulting" is therefore a subset of the practice list of all registered patients. Consultations must be carefully distinguished from "patients consulting." A combination of patient number, date and place of consultation and diagnosis uniquely define each record in the consultation file. Patient numbers are not unique because a patient may consult more than once, nor are combinations of patient number and diagnosis unique. On the other hand, a "patient consulting" file will contain at most one record for each patient consulting for a particular diagnosis (or group of diagnoses), no matter how many times that patient has consulted a member of the practice staff. "Consultations" are more relevant when work-load is being studied, but if prevalence is the issue then "patients consulting," i.e., how many patients consulted for the illness, is more useful.

- **Patient Years at Risk.--**The population involved in the MSGP project did not remain constant throughout the study. Patients entered and left practices as a result of moving house or for other reasons, and births and deaths also contributed to a changing population. The "patient years at risk" derived variable was created to take account of this. The patient file contains a "days in" variable, which gives the number of days the patient was registered with the practice (range 1-366 days for the study). "Patient years at risk" is "days in" divided by 366, since 1992 was a leap year.

Database Structure

To facilitate future analyses some non-changing data were combined at the outset. For example some consultation metadata were added to the consultation dataset, such as International Classification of Disease (ICD) codes and indicators of disease seriousness. The resultant simplified data structure is thus:

Practice: Practice number; information about practice (confidential)
Primary Key: Practice number
A practice is a group of doctors, nurses, and other staff working together. Although patients register with a particular doctor, their records are kept by the practice and the patient may be regarded as belonging to a practice. Data on practice and practice staff are particularly confidential, and not considered in this paper. Individual practice staff consulted are identified in the consultation file by a code.

Patients: Patient number; age; sex; post code; socio-economic data
Primary key: Patient number
Foreign key: Postcode references geographic data
These data were stored as four separate files relating to: all patients; adult patients; children; married cohabiting women, because different information was collected for each subgroup.

Consultation: Patient number; Practice number; ID of who consulted; date of contact; diagnosis; place of consultation; whether referred to hospital; other consultation information
Primary key: Patient number, doctor ID, date of contact, diagnosis
Foreign keys: Practice number references practice; Patient number references patients; Staff ID references staff (e.g., doctor/nurse).

Episode: For each consultation the doctor/nurse identified whether this was the "first ever," a "new," or "ongoing" consultation for that problem. An "episode" consists of a series of consultations for the same problem (e.g., Read code).

Geographically-referenced data: Post codes, ED, latitude/longitude, census ward, local authority, small area census data, locality classifications such as rural/ urban, prosperous/inner city, etc.

These data were not collected by the survey, but come from other sources, linked by postcode or higher level geography.

Patient metadata: These describe the codes used in the socio-economic survey (e.g., ethnic group, occupation groups, social class, housing tenure, whether a smoker, etc.)

Consultation metadata: The ReadICD file links Read codes with the corresponding ICD codes. In addition a lookup table links 150 common diseases, immunisations and accidents to their ICD codes. Each diagnosis is classified as serious, intermediate or minor.

Derived files: The MSGP database contains information on individual patients and consultations. To make comparisons between groups of patients, and to standardise the data (e.g., for age differences), it is necessary to generate files of derived data, using database queries and linkages as described below. In some derived files duplicate records need to be eliminated. For example, we may wish to count **patients** consulting for a particular reason rather than consultations, and hence wish to produce at most one record per patient in a "patients consulting" derived file -- see "Some Definitions above).

Types of Linkage (with Examples)

In this section we classify a variety of linkage types that are possible into three main types, illustrating the linkages with examples based on the MSGP4 study.

Simple Linkage

- Straightforward data extracts (lists) combining several sources.—

  Example: Making a list of patients with asthma including age, sex and social class for each patient.

- Observed frequencies.—

  Example: Linking the "all patients" file, and the "consultations" file to count the number of consultations by the age, sex and social class of the patient, or cross-classifying home-visits and hospital referrals with socio-economic characteristics.

- Conditional data, where the availability of data items depends on the value of another variable.—

  Example: In MSGP4 some data are available only for adults, or children, or married/cohabiting women. Smoking status was only obtained from adult patients, so tabulating "home visits" by "smoking status" by "age," and "sex" involves linking the "all patients" (to find age and sex), "adult patients" (to find smoking status) and "consultations" (to find home visits) files. Linking the "adult" file to the "all patients" file excludes records for children.

- Linking files with "foreign" files.— Useful information can often be obtained by linking data in two or more different datasets, where the data files share common codes. For example data referenced by postcode, census ED or ward, or local authority are available from many different sources as described above.

  Example: The MSGP4 study included the postcode of residence for each patient, facilitating studies of neighbourhood effects. The crow-fly distance from the patient's home to the practice was calculated by linking patient and practice postcodes to a grid co-ordinates file and using Pythagoras's theorem. The distance was stored permanently on the patient file for future use.
- Linking to lookup tables (user-defined and pre-defined).—

  Examples: The information in the MSGP database is mostly held in coded form, with the keys to the codes held in a number of lookup tables linked to the main database. Most of these are quite small and simple (e.g., ethnic group, housing tenure, etc.) but some variables are linked to large tables of standard codes (e.g., occupational codes, country of birth). . In some cases the coded information is quite detailed and it is desirable to group the data into broader categories, e.g., group diagnostic codes into broad diagnostic groups such as ischaemic heart disease ICD 410-414. For some diseases a group of not necessarily contiguous codes are needed to define a medical condition. A lookup file of these codes can be created to extract the codes of interest from the main data, using a lookup table that could be user-defined. Missing value codes could also be grouped, ages grouped into broad age groups, social classes combined, etc.

Auto-Linkage Within a File (Associations Within a File)

- Different records for the same "individual."— Records for the same individual can be linked together to analyse patterns or sums of events, or associations between events of different kinds. In general a file is linked to a subset of itself to find records relating to individuals of interest.

  Example: Diabetes is a chronic disease with major complications. It is of interest to examine, for those patients who consulted for diabetes, what other diseases they consulted for. Consultations for diabetes can be found from their ICD code (250). Extracting just the patient identification numbers from this dataset, and eliminating duplicates, results in a list of patients who consulted for diabetes at least once during the year. This subset of the consultation file can be linked with the original consultation file to produce a derived file containing the consultation history of all diabetic patients in the study, which can be used for further analysis. Note that in this example only the consultation file (and derived subsets) has been used.

- Different records for same households/ other groups.—

  Example: Information on households was not collected as part of MSGP4. However "synthetic" households can be constructed, using postcode and socio-economic data, where the members of the same "household" must, by definition, share the same socio-economic characteristics and it would be rare for two distinct households to have exactly the same characteristics. These "households" can be used to discover how the behaviour of one "household" member may affect another. For example, we can examine the relationship between smoking by adults, and asthma in children. Clearly in this example some sort of check needs to be made on how accurately "households" can be assembled from the information available and the algorithm used.

- Temporal relationships.— Files containing "event" data can be analysed by studying temporal patterns relating to the same individual.
  Example:
  - The relationship between exposure to pollution or infection and asthma can be studied in terms of both immediate and delayed effects. Consultations for an individual can be linked together and sorted by date, showing temporal relationships.

  - The duration of clinical events can sometimes be determined by the sequence of consultations. In MSGP4 each consultation for a particular medical condition was labelled "first ever," "new," or "ongoing" and the date of each consultation recorded. Survival analysis techniques cater for these types of data.

Complex Linkages

Linkages that are combinations of the two types of linkage previously described could be termed "complex linkages." These can always be broken down into a sequence of simpler linkages. A number of examples of complex linkages are given, in order of complexity.

- Finding subsets through linkage.—

  Example: In the MSGP4 data this is particularly useful in the study of chronic conditions such as diabetes and heart disease. Linking the file of patients consulting for diabetes discussed in section 3.2 with the patient dataset results in a subset of the patient file, containing only socio-economic details of diabetic patients.

- Linking a derived file to a lookup table and other files.—

  Example: Diabetes is particularly associated with diseases of the eye (retinopathy), kidney, nervous system and cardiovascular systems. It is of interest to analyse the relationship between diabetes and such diseases, which are likely to be related to diabetes. In this slightly more complex situation it is necessary to create a lookup table containing the diseases of interest and their ICD codes and link this to the "consultations by diabetic patients" file to create a further subset of the consultation file containing consultations for diabetes and its complications. It is likely that this file as well as the simpler one described above would be linked to the patient file to include age and sex and other patient characteristics before analysis using conventional statistical packages.

- Linking a derived file with another derived file.—

  Example:
  - Rates for groups of individuals.— Rates are found by linking a derived file of numerators with a derived file of denominators. The numerators are usually found by linking the patient and consultation files, for example, age, sex, social class or ethnic group linked to diagnosis, referral or home visits. Denominators can be derived from the patient file (patient years at risk) or the consultation file (consultations or patients consulting) for the various categories age, sex, etc.

  - Standardised ratios.— This is the ratio of the number of events (e.g., consultations or deaths) observed in a sub-group to the number that would be expected if the sub-group had the same age-sex-specific rates as a standard population (e.g., the whole sample), multiplied by 100. Examples of sub-groups are different ethnic groups or geographical areas. The calculation of standard population rates involves linking the whole population observed frequencies to whole population patient years at risk. Each of these is a derived file, and the result is a new derived file. Calculating expected numbers involves linking standard population rates to the sub-groups' "years at risk" file. This produces two new derived files, "Observed" and "Expected." Age-standardised patient consulting ratios are obtained by linking these two derived files together, using outer joins to ensure no loss of "expected" records where there are no observed in some age-sex categories.

- Establishing population rates for a series of nested definitions.—

  Example: Individuals at particular risk from influenza are offered vaccination. In order to estimate how changes in the recommendations might affect the numbers eligible for vaccination, population rates for those living in their own homes were estimated for each of several options. People aged 65 and over living in communal establishments are automatically eligible for vaccination, and hence were selected out and treated separately. The options tested were to include patients with:

  A - any chronic respiratory disease, chronic heart disease, endocrine disease, or immune-suppression;

B - as A but also including hereditary degenerative diseases;
C - as B but also including thyroid disease;
D - as C but also including essential hypertension.

The MSGP dataset was used to estimate the proportion of the population in need of vaccination against influenza according to each option. The problem was to find all those patients who had consulted for any of the diseases on the list, taking care not to count any patient more than once. This involved creating a lookup table defining the disease groups mentioned in options A-D, linking this to the consultation dataset, eliminating duplicates and linking this to the patient dataset (to obtain age-group and sex), and then doing a series of queries to obtain appropriate numerator data files. A denominator data file was separately obtained from the patient dataset to obtain patient years at risk, by age-group and sex. The numerator and denominator files were then joined to obtain rates. These rates were then applied to census tables to obtain the estimated numbers of patients eligible for vaccination under assumptions A-D.

- Record matching for case-control studies.— These are special studies of association-extracting "cases" and "controls" from the same database.

  Example: what socio-economic factors are associated with increased risk of Crohn's disease? All patients who consulted for ICD555 (regional non-infective enteritis) during the MSGP4 study were selected and referred back to their GP to confirm that they were genuine cases of Crohn's disease. Patients who were not confirmed as having Crohn's disease were then excluded. This resulted in 294 cases. Controls were selected from patients who did have the disease – those who matched cases for practice, sex and month and year of birth. In each of two practices there were two cases who were of the same sex and the same month and year of birth. In each of these practices the controls were divided randomly between these cases as equally as possible. There were 23 cases for whom no controls could be found using these criteria. In 20 of these cases it was possible to find controls who matched on practice and sex and whose date of birth was within two months of the case's date of birth. The remaining three cases were excluded from the analysis. This procedure resulted in 291 cases and 1682 controls.

User-Friendly Linkage Software

The MSGP4 practice software was originally written so that participating practices could gain access to the data collected from their own practice. The software was designed to be used easily by people with no knowledge of database technology and because the software runs directly under DOS or Windows, no specialised database software is needed. The structure of the MSGP database is transparent to the user who can refer to entities (e.g., diseases or occupation) by name rather than codes.

Later, a modified version of the software was developed to enable researchers to use the complete dataset (60 practices).

Although it may be possible for some of these linkages to be performed as a single query it is generally best to do a series of simple linkages for two reasons. Firstly, database software creates large temporary files of cross products, which is time consuming and may lead to memory problems. Secondly, queries involving complex linkages are often difficult to formulate and may easily turn out to be incorrect. The order in which the linkages are performed is also important for efficiency. In general, only the smallest possible files should be linked together. For example, rather than linking the patient and consultations files together, then finding the diseases and patient characteristics of interest, it is better to find the relevant subsets of the two files first, then link them together.

The software performs the required linkages and then analyses the data in two stages. The first part of the program performs the sequence of linkages and queries needed to find subsets required for the second stage, and the second part performs the analyses and displays the output. The data flow through the program is shown in Figure 3.

It can be seen from the diagram that any of the three input files may be linked to themselves or to either of the others in any combination to form subsets of the data, or the entire dataset can be used.

Finding Subsets

- The program enables the user to find any combination of characteristics required, simply by choosing the characteristic from menus. The program finds subsets of individual files, as well as linking files in the dataset to each other and to lookup tables, and finding subsets of one file according to data in another. For example the program can produce a list of young women with asthma who live in local authority accommodation, or of patients with a particular combination of diagnoses. It is also possible to examine the data for a particular group of people (for example, one ethnic group), or for a particular geographical area.

- Dealing with missing values.--When the data for MSGP4 was collected it was not possible to collect socio-economic data for all patients. The user is given the option to exclude missing values, or to restrict the data to missing values only should they want to find out more about those patients for whom certain information is missing. For example, an analysis of the frequency of cigarette smoking in each age/sex group in the practice might include only those patients for whom smoking information is available.

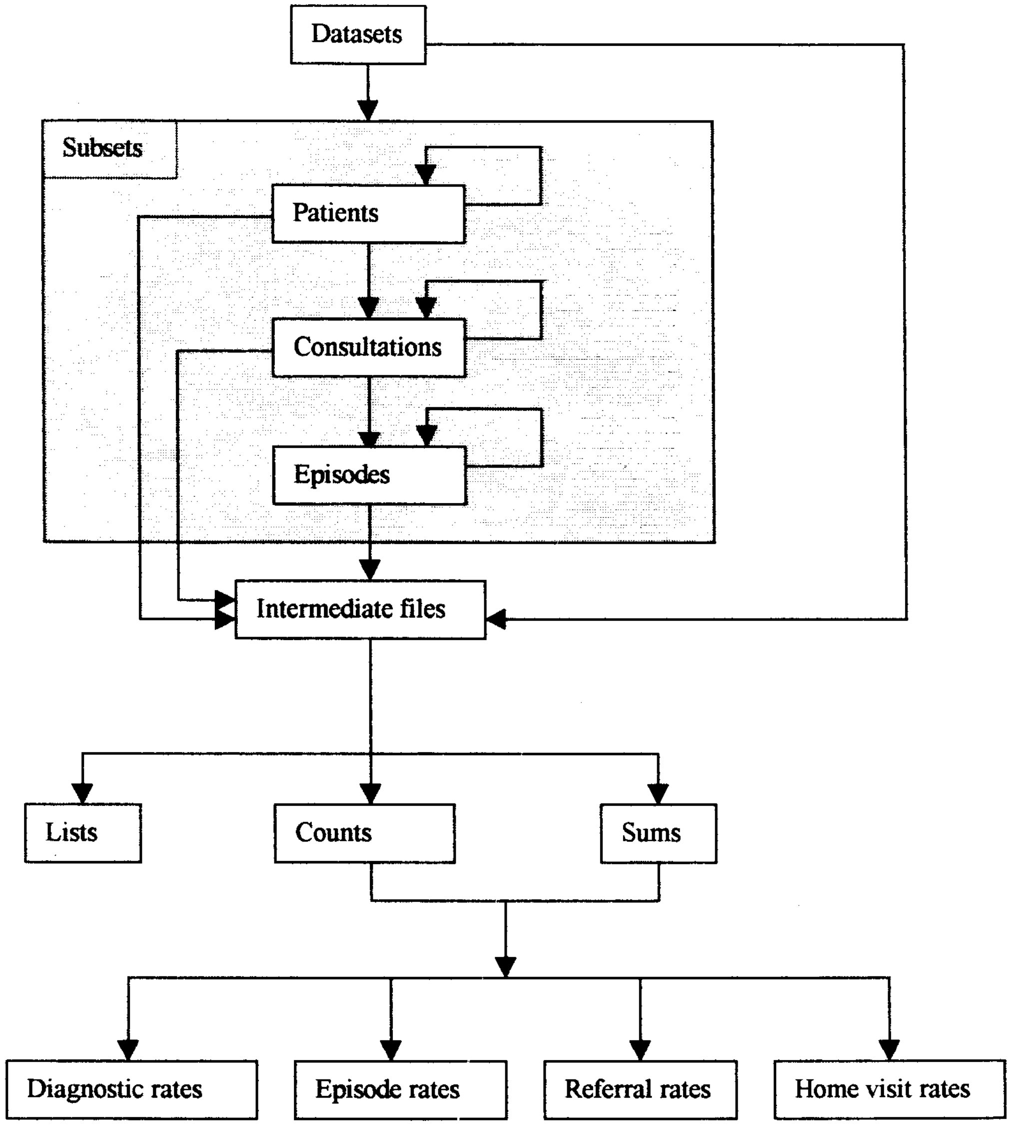
The Output

The output from the program is of three types, any of which may be exported by the program in a variety of formats (e.g., WK1, DBF, TXT, DB) for further statistical analyses.

- Lists output consists of one record for each patients, consultation or episode of interest, with files linked together as appropriate. Each record contains a patient number together with any other information that the user has requested. These flat files can be used for further analysis using spreadsheet or statistical software.

- Frequency output consists of counts of the numbers of patients, consultations or episodes in each of the categories defined by the fields selected by the user.
- Rate output enables a variety of rate with different types of numerators and denominators to be calculated. Any of the following rates may be chosen: Diagnostic rates for a specified diagnostic group (patients consulting; consultations; episodes); referral rates; and home visit rates. Rates are generally calculated for standard age and sex groups but other appropriate patient and consultations characteristics may be included in the analysis. Denominators can be consultations, patients consulting or patient years at risk.

Figure 3.—Data-flow Diagram for MSGPX Data Extractor Program


Datasets
Subsets
Patients
Consultations
Episodes
Intermediate files
Lists
Counts
Sums
Diagnostic rates
Episode rates
Referral rates
Home visit rates

Discussion and Conclusions

We have demonstrated through the use of one example database the potential that relational databases offer for storing statistical data. These are also the natural way to capture the data, since they reflect real data relationships, and are economical in storage requirements. They also facilitate linking in new data from other sources. However most statistical analyses require simple rectangular files, and complex database queries may be required to obtain these. We have shown that such complex linkages can be decomposed into a sequence of simple linkages, and user-friendly software can be developed to make such complex data readily available to users who may not understand the data structure or relational databases fully. The major advantage of such software is that the naïve user can be more confident in the results than if they were to extract the data themselves. They can also describe their problem in terms closer to natural language.

Although such programs enable the user with no knowledge of database technology to perform all the linkages shown above, they do have their limitations. Choosing options from several dialogue boxes is simple but certainly much slower than performing queries directly using SQL, Paradox or other database technology. Since the most efficient way to perform a complex query depends on the exact nature of the query, the program will not always perform queries in the most efficient order. The user is also restricted to the queries and tables defined by the program, and as more options are added the program must of necessity become more unwieldy and possibly less efficient.

User friendly software remains, however, the useful for the casual user who may not be familiar with the structures of a database, and essential for the user who does not have access to or knowledge of database technology.

References

Fellegi, I.P. and Sunter, A.B. (1969). A Theory for Record Linkage, *Journal of the American Statistical Association*, 64: 1183-1210.

McCormick, A.; Fleming, D.; and Charlton, J. (1995). *Morbidity Statistics from General Practice*, Fourth National Study 1991-92, Series MB5, no 3, London: HMSO.

Newcombe, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*, Oxford: Oxford University Press.

Newcombe, H. B.; Kennedy, J. M.; Axford, A. P.; and James, A. P. (1959). Automatic Linkage of Vital Records, *Science*, 130: 954-959.

Winkler, W. E. (1994). Advanced Methods of Record Linkage, *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 467-472.